DOCUMENT RESUME

ED 408 329 TM 026 582

AUTHOR Wise, Steven L.
TITLE Examinee Issues in CAT.
PUB DATE Mar 97
NOTE 17p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, March 25-27, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; Computer Attitudes; Equal Education; Item Banks; Review (Reexamination); *Student Attitudes; Student Motivation; Test Anxiety; *Test Construction; Test Items; *Testing Problems; Timed Tests
IDENTIFIERS Calibration

ABSTRACT

The perspective of the examinee during the administration of a computerized adaptive test (CAT) is discussed, focusing on issues of test development. Item review is the first issue discussed. Virtually no CATs provide the opportunity for the examinee to go back and review, and possibly change, answers. There are arguments on either side of the item review issue, and test givers should weigh them carefully, considering examinee anxiety and performance factors. Another issue is that of time limits, which have little benefit for test takers, but serve only the interests of test givers. CAT developers should consider very liberal time limits or none at all, especially since a CAT is shorter than its conventional testing counterparts. Test anxiety may be increased in a CAT environment, and test developers should be aware of the potential for anxiety among examinees. Another issue is that of examinee motivation. CAT developers should be aware of the effects of test consequences on test performance to ensure that data used to calibrate item banks are collected under conditions that have the same consequences as the operational test. Equity is an important issue in CAT, since some examinees will have less computer experience than others. Each of these issues has implications for the validity of inferences made from CAT scores and should be considered when CATs are used. (Contains 27 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Steven L. Wise
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 408 329

Examinee Issues in CAT

Steven L. Wise, University of Nebraska-Lincoln

Paper presented at the 1997 annual meeting of the National Council on Measurement in Education, Chicago, IL

TM026582

BEST COPY AVAILABLE

Examinee Issues in CAT

The computerized adaptive test (CAT) is rapidly becoming a familiar mode of test administration. Many large-scale testing programs are implementing CATs, either as an alternative to conventional multiple-choice test versions (e.g., the Graduate Record Exam), or as the only available testing format. This trend is likely to continue—or perhaps increase—as more testing programs seek ways to efficiently administer tests that are often quite lengthy in their conventional (i.e., fixed paper-and-pencil) forms.

The mathematical methods and computer algorithms used in operational CATs are very straightforward and compelling in their appeal. Combining the advantages of item response theory (IRT) over classical test theory with the computing power of current microcomputers, a CAT promises to (a) efficiently measure examinee proficiency and (b) provide immediate test performance feedback to examinees. As such, a CAT represents a unique, practical contribution to modern measurement. Its efficiency, moreover, is extremely attractive in a society that already gives many tests—and appears predisposed toward more, not less, testing in the future.

Although the idea of adaptive testing is conceptually simple, the development and maintenance of a CAT program is much more complicated. As the other papers in this symposium have discussed, the test administrator (i.e., the test giver) must find adequate solutions to a number of practical technical problems. A CAT is about testing people, however, and test givers would be prudent to not overlook potential problems that a CAT administration might cause for examinees. Although it is important to consider the perspective of the examinee in any proficiency measurement, it is particularly important for us to

understand how the unique—and relatively new—testing methods used in a CAT affect examinees.

At first glance, the experience of taking a CAT may not appear to be very different from a conventional test. An item appears on the computer screen, the examinee develops and enters his or her answer, and the next item appears. This process continues until the test is completed.

There are, however, a number of unique aspects to the CAT experience that might influence an examinee's test performance. First, computer-based test administration may not be a familiar mode of testing to many examinees. Items presented on a computer screen may be more difficult or fatiguing for examinees to read, particularly longer items whose size exceeds the computer screen and that require examinees to scroll through the item content. The entry of examinee responses using a keyboard is different from circling an answer on a test booklet or filling in a bubble on a machine-scorable answer sheet.

Second, in a conventional test, examinees are given all of their test items at once. This provides examinees a great deal of freedom regarding browsing through all the items, skipping some to be answered at the end of the test, and reviewing—and possibly changing—answers. In contrast, examinees have far less control when taking a CAT.

Third, all but the most naive examinees will have some idea that there is some sort of computer algorithm operating that is used in identifying which items are administered. That is, examinees have a sense that they are interacting with the computer, and that how they behave—in terms of test performance—affects what how the computer behaves. The presence of this interaction may have an effect on examinees.

Finally, in many CATs the length of the test (in number of items) can vary markedly across examinees. In norm-referenced measurement, different test

lengths result whenever a common standard error of proficiency estimation is used as the criterion for terminating the CATs of all examinees. In a criterion-referenced measurement context in which the goal of measurement is to identify examinees whose proficiency levels exceed some standard, testing for a given examinee will continue only until a confident pass/fail decision can be made. During these types of testing situations, examinees will have little idea how close they are to the end of their tests. This is quite different from conventional tests, in which examinees can continually tell how close they are to completing their tests, and budget their efforts accordingly.

The purpose of this paper is to discuss the examinee's perspective during a CAT administration. Five examinee issues will be discussed. The issues are inter-related, as decisions made by the test giver concerning each issue may affect other issues as well. As part of the discussion for each issue, relevant research findings will be presented, and recommendations for practitioners will be given.

Item Review

Although the reaction of examinees to CATs has been generally positive, they have expressed one major concern. Virtually no operational CATs provide an opportunity for examinees to go back and review—and possibly change—any of their answers to previously administered items. Research has consistently reported that examinees express dissatisfaction with the lack of item review(Baghi, Ferrara & Gabrys, 1992; Legg & Buhr, 1992; Vispoel, Rocklin & Wang, 1994; Vispoel, Wang, Torre, Bleiler & Dings, 1992). Anecdotally, when I informally ask graduate students about their GRE-CAT experiences, they often state being bothered by the absence of review. And in our own CAT testing at the University of Nebraska, the most frequently asked question by examinees as they are being given their tests is, "Will I be able to go back and check my

answers?". Clearly, then, examinees attend to whether item review is provided, and many are bothered when it is not.

Why should test givers be concerned about this? The availability of item review to examinees during conventional tests was an unplanned, uncontrollable consequence of the development of group-administered achievement and ability tests. With computer-based testing, however, test givers can effectively prevent examinees from reviewing their answers. Moreover, if everyone is denied item review on a CAT, then everyone is treated the same. This test giver-imposed control over item review is therefore consistent with test standardization.

Denying item review may, however, may have negative consequences for examinee test performance. Over sixty years of research has consistently shown that (a) when examinees are allowed to change answers, they are more likely to improve their scores, and (b) score gains due to answer changes are overwhelmingly due to legitimate reasons, such as rethinking or rereading the item, or making a clerical error. A recent paper of mine (Wise, 1996) overviews much of the research in this area and provides a discussion of the relevant issues within the context of CATs (although most of the arguments apply to non adaptive tests as well). It follows that denying item review denies an opportunity for answer changing, which tends to improve scores.

There is also the possibility that denying item review results in increased levels of anxiety—and possibly impaired test performance—for some examinees. While denying item review represents increased control for the test giver, it also means decreased control for the examinee. And it has been found, in many contexts, that individuals better tolerate stressful situations (such as tests) when they feel that they have some control over their environment. Increased perceived control has been associated with decreased anxiety and improved task performance. See Wise (1994) for an overview of this research.

The effects of increased perceived control are often moderated, however, by an individual's desire for control (Burger, 1989). That is, positive effects of increased perceived control are observed only for individuals who desire such control. Moreover, it has been found that examinees vary substantially in their desire for control in an examination context (Wise, Roos, Leland, Oats & McCrann, 1996). This all suggests that any decreases in perceived control associated with the denial of item review may not affect all examinees equally—which would imply that denying item review for all examinees would have a differential effect across examinees. While this argument is largely speculative, it raises an issue that may affect the validity of inferences made from CAT scores when item review is denied.

Recommendation

There are arguments on either side of the item review issue (Wise, 1996), and test givers should carefully weigh these arguments in deciding whether item review should be provided. It is important to consider the examinee's perspective in making these decisions.

Time Limits

Placing a limit on the time that examinees may spend on a test or test section is a typical feature of standardized tests. Time limits, however, serve only the interests of the test givers, who are motivated to administer a test as efficiently and cheaply as possible. For an examinee's perspective, time limits have little benefit; on the contrary, time limits add to the stress of the testing context, and undoubtedly increase the anxiety levels for many examinees.

Establishing a reasonable time limit for a test is a tricky business. If the testing time is too long, then time needed to administer a test is needlessly lengthened, with consequent loss in time and money. If the testing time is too short, then some examinees will not be able to complete all of the test items in the

allotted time. For these examinees, the resultant test scores will underestimate their true levels of proficiency—which means that the test validity has been compromised.

For a CAT, however, establishing a time limit is more complicated. One reason is that CATs that use score reliability as a stopping criterion will administer tests of different lengths. And if one does not know in advance how long a given examinee's test will be, how does one know how much time to allow? Even when fixed-length CATs are used, the time limits issue is complex. Imagine two CAT examinees: a more able examinee who receives 40 harder math items, and a less able examinee who receives 40 easier math items. Should the same time limit be used? What if it were known that the harder items generally required more time for an examinee to answer, because they involved more time-consuming computations? Because examinees each receive a unique set of test items, it is more difficult to choose a single time limit that would be appropriate for each of these tests.

The issue of appropriate time limits to provide on a CAT is a challenging issue. Indeed, one might argue that the imposition of a time limit is antithetical to a goal of a testing program that promotes students exhibiting their optimal levels of performance. The goal is to identify a time limit that does not meaningfully limit student performance, while keeping the testing session reasonably short. This issue is complicated by research indicating that some ethnic minority groups take more time to complete CATs (Baghi et al., 1992; Legg & Buhr, 1992; O'Neill & Powers, 1993; Zara, 1992), although some research has indicated that allowing minority students more time on conventional tests has not enhanced their performance relative to majority students (Evans & Reilly, 1972; Wild, Durso & Rubin, 1982).

The relationship between time limits and test performance appears to be moderated by examinee test anxiety. Research has shown that the differences in test performance between timed and untimed tests are greater for highly test anxious examinees than for examinees reporting less anxiety (Hill, 1984; Onwuegbuzie & Seaman, 1995). This suggests that lengthening a time limit on a CAT may help some examinees more than others. Or, put another way, a time limit that is too short may have a greater impact on test anxious examinees.

Recommendation

Given the differences among examinees, it appears that a single time limit is likely to be difficult to defend as equitable. Therefore CAT developers should adopting very liberal time limits, or consider imposing no time limits at all. Keep in mind that a CAT is dramatically shorter than its conventional counterpart; we should consider giving some of that saved time back to examinees. Examination-related stress would thereby be reduced and test validity may be enhanced.

Test Anxiety

The relationship between anxiety and test performance has been extensively studied. Although a number of theories of test anxiety have been proposed, it has generally been found that increased anxiety leads to decreased performance (Hembree, 1988; Schwarzer, Seipp & Schwarzer, 1989). It has been estimated that up to 10 million U.S. students are affected significantly each year by the debilitating effects of test anxiety (Hill, 1984).

Although examinees vary in their tendencies to become anxious during tests, the anxiety experienced in a particular instance of testing is a function of both the tendency of the examinee to experience anxiety and the setting and manner in which the test is administered. Thus, felt anxiety during a test has both state and trait components, and can be altered (to some extent) by the

testing environment. Are there unique aspects of a CAT, relative to a conventional test, that might increase anxiety in some or all examinees?

An obvious candidate is the computer itself. Does examinee computer anxiety or inexperience with computers interfere with student test performance? Computer anxiety and experience are considered together because (a) they have frequently been studied together in the research literature and (b) they appear to show a strong inverse relationship (greater experience is associated with less anxiety). Research has generally shown computer anxiety and experience to be unrelated to test performance (Kim & McLean, 1994; Powers & O'Neill, 1992; Wise, Barnes, Harvey & Plake, 1989), although Legg and Buhr (1992) found anxiety during a CAT to be inversely related to computer experience.

There are, however, several other potential sources of anxiety in a CAT environment. First, as discussed earlier, the absence of item review may be anxiety provoking. Second, an examinee can often sense whether his or her items are getting easier or harder. Easier items mean poor test performance, which—if noticed by the examinee—can increase anxiety felt during the test. Third, because examinees will often know that a CAT typically results in substantially shorter tests being administered, they may infer that each item now has a larger impact on their final scores. With more riding on each item, examinees may feel more stress and greater anxiety. Finally, the items received on a CAT are much more homogeneous in difficulty than a conventional test. Moreover, the proportion of items passed on a CAT is typically far lower than examinees are used to experiencing with a conventional test. These differences also hold potential for increasing the anxiety levels of some examinees, because they will perceive a diminished feeling of mastery over the test items.

The research on the effects of CATs on examinee anxiety has yielded mixed results. Legg and Buhr (1992) found that a feeling of anxiety in a CAT

testing situation varied across examinee gender, ethnic, and ability groups. Baghi et al. (1992), however, found no differences in anxiety among gender and ethnic groups, and found an anxiety difference across ability groups for only one of the two CATs that they studied. Furthermore, there is evidence that the relationship between test anxiety and performance is weaker for a CAT than for a conventional test (Gershon & Bergstrom, 1991).

Recommendation

The field of educational measurement needs to better understand the impact of a CAT on examinee anxiety and performance. We also should not be very satisfied to observe a lack of overall mean differences in anxiety and performance between groups of examinees testing under CAT and conventional conditions. Small mean differences may obscure a situation in which a CAT is meaningfully affecting the anxiety levels of a relatively small proportion of the examinees. It is important that these types of situations be identified, and that corrective action be taken for these examinees (e.g., providing a conventional test).

Examinee Motivation

Another examinee variable that is related to test performance is motivation. That is, if examinees are not motivated to do their best, test performance will be adversely affected. Examinee motivation and consequent test performance have been shown to be influenced by the perceived consequences associated with test performance (Brown & Walberg, 1993; Kim & McLean, 1995; Wolf & Smith, 1995).

This relationship between motivation and test performance is relevant to a testing program in which an IRT-calibrated item pool is developed and maintained (such as with a CAT). Establishing an item pool, which typically contains hundreds of items, requires a lot of data. Depending on the IRT model

used, the minimum recommended numbers of examinees taking each item ranges from 200 up toward 1000. This is quite a logistical challenge to testing programs.

Many CAT programs evolve from established conventional testing programs. This typically means that there is a lot of items from previously used test forms (with accompanying data) that could be used in the item bank. A prudent test developer, however, would recognize that the item parameters for a paper-and-pencil version of an item may not be the same as those for a computer-based version. A safer solution would be to (a) develop a set of fixed computer-based tests that collectively contain all of the items in the pool, (b) computer administer them to a sufficient number of examinees, and (c) base the item calibrations on those data.

If, however, these fixed tests are administered under nonconsequential conditions—either as practice tests or given to groups of volunteers—then the item calibrations are likely to be biased. The examinees will not be as motivated—which means that they collectively will not do as well on the items. The result will be negatively biased difficulty parameters, because the items will appear to be more difficult than they would under consequential testing conditions. And once this item bank is subsequently used by an operational CAT, examinee scores will be positively biased, because examinees will appear to be passing more difficult items. This effect is similar to that noted by Wolf and Smith (1995) that test norms established under nonconsequential conditions may lead to inflated norm-referenced performance by future examinees under consequential conditions.

Recommendation

CAT developers should be aware of the effects of test consequences on test performance. They should ensure that the data used to calibrate item banks are

collected under conditions that have the same consequences as will be observed during an operational test.

Equity

Sutton (1993) posed an important question that is relevant to CAT: Will the use of computer-based testing maintain or exaggerate inequalities in education? Sutton also raised the related issue of how computer-based testing can be used to reduce inequities. Regarding differences among gender and racial/ethnic groups, is a CAT likely to make things better or worse? These questions will be discussed in terms of several of the examinee issues previously discussed.

There is evidence that poor and minority children have had less access to computers at home and at school (Sutton, 1993). Because less access implies less experience, the relationship between computer experience and CAT performance becomes of increased importance. Research on this issue specifically related to CATs is mixed. One study found differences among racial/ethnic groups (Buhr & Legg, 1989a) on computer usage, while the other (Baghi et al., 1992) did not.

As discussed earlier, there are differences in racial/ethnic groups concerning testing time used on a CAT (Baghi et al., 1992; Legg & Buhr, 1992; O'Neill & Powers, 1993; Zara, 1992). Hence, any time limit that is imposed may have a differential effect in different groups—which may exacerbate test performance differences among these groups.

What has the research shown regarding subgroup differences in test performance between computer-based and conventional tests? Johnson and Mihal (1973) compared the performance of Black and White examinees on computer-based and conventional fixed-item forms of the School and College Ability Tests, finding that computer administration resulted in higher scores for Blacks but not for Whites. Research regarding the effects of CATs is mixed. Zara

(1992) found that the differences in performance between computer-based and conventional versions of a national nursing licensure exam varied substantially across ethnic groups. White examinees showed a modest difference in favor of the conventional version, whereas Black examinees showed virtually no difference in performance between the test versions. In contrast, Buhr and Legg (1989b) found that, although all ethnic groups scored higher on their CAT reading test, differences between scores for White examinees and those for Blacks and Hispanics were greater on the conventional test than on the CAT. Hence, the limited research regarding subgroup differences in test performance between CAT and conventional tests has not indicated that ethnic minority groups would be disadvantaged by a CAT.

Recommendation

At this point, it is too early to tell whether use of a CAT is likely to increase or decrease test score differences among subgroups. Test developers should, however, be prepared to investigate this issue with their own CATs. Again, adopting liberal time limits is likely to minimize any subgroup score differences that are attributable to differences in the time needed to take a CAT.

Conclusions

In this paper, I have attempted to identify and discuss five examinee issues that should be considered by developers of CATs. Each of these issues has implications for the validity of inferences made from CAT scores. And because test developers have a responsibility to promote test score validity for all examinees, it is crucial that these examinee issues be given attention when developing a CAT.

The mechanics of a CAT are well understood. We know far less, however, about how CATs affect examinees. We should not be content to simply randomly assign a group of examinees to conventional and CAT testing

conditions, and if the groups' mean test scores do not differ significantly conclude that the testing formats are equivalent. Issues such as time limits, anxiety, or a lack of item review may impact only a small proportion of the examinee population. Being in the minority, however, does not mean unimportant. The needs of all examinees are important, and we must consider all relevant influences on examinees when CATs are used. It is through a better understanding of the psychological dynamics underlying test taking that we will be able to fully understand which dynamics are important to examine in developing a CAT.

References

Baghi, H., Ferrara, S. F., & Gabrys, R. (1992, April). Student attitudes toward computer-adaptive test administrations. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Brown, S. M., & Walberg, H. J. (1993). Motivational effects on test scores of elementary students. Journal of Educational Research, 86, 133-136.

Buhr, D. C., & Legg, S. M. (1989a, March). Investigating the validity of a computerized adaptive test for different examinee groups. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Buhr, D. C., & Legg, S. M. (1989b). Report on the study of the computerized adaptive CLAST reading, writing, and mathematics tests (Contract: 88012704). University of Florida: Institute for Student Assessment and Evaluation.

Burger, J. M. (1989). Negative reactions to increases in perceived personal control. Journal of Personality and Social Psychology, 56, 246-256.

Evans, F. R., & Reilly, R. R. (1972). A study of test speededness as a potential source of bias in the admission test for graduate study in business quantitative score. . Princeton, NJ: Educational Testing Service.

Gershon, R. C., & Bergstrom, B. (1991, April). Individual differences in computer adaptive testing: Anxiety, computer literacy and satisfaction. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. Review of Educational Research, 58, 47-77.

Hill, K. T. (1984). Debilitating motivation and testing: A major educational problem—Possible solutions and policy implications. In R. E. Ames & C. Ames (Eds.), Research on motivation in education (Vol. 1, pp. 245-274). New York: Academic Press.

Johnson, D. F., & Mihal, W. L. (1973). Performance of blacks and whites in computerized versus manual testing environments. American Psychologist, 28, 694-699.

Kim, J., & McLean, J. E. (1994, November). The relationships between individual difference variables and test performance in computerized adaptive testing. Paper presented at the annual meeting of the Mid-South Educational Research Association, Nashville, TN.

Kim, J., & McLean, J. E. (1995, April). The influence of examinee test-taking motivation in computerized adaptive testing. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Legg, S. M., & Buhr, D. C. (1992). Computerized adaptive testing with different groups. Education Measurement: Issues and Practice, 11, 23-27.

O'Neill, K., & Powers, D. E. (1993, April). The performance of examinee subgroups on a computer-administered test of basic academic skills. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Onwuegbuzie, A. J., & Seaman, M. A. (1995). The effect of time constraints and statistics test anxiety on test performance in a statistics course. Journal of Experimental Education, 63, 115-124.

Powers, D. E., & O'Neill, K. (1992). Inexperienced and anxious computer users: Coping with a computer-administered test of academic skills. The Praxis Series: Professional Assessments for Beginning Teachers. (RR-92-75). Princeton, NJ: Educational Testing Service.

Schwarzer, R., Seipp, B., & Schwarzer, C. (1989). Mathematics performance and anxiety: A meta-analysis. In R. Schwarzer, H. M. van der Ploeg, & C. D. Spielberger (Eds.), Advances in test anxiety research (Vol. 6, pp. 105-120). Lisse, The Netherlands: Swets & Zeitlinger.

Sutton, R. E. (1993, April). Equity issues in high stakes computerized testing. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized adaptive, and self-adapted testing. Applied Measurement in Education, 7, 53-59.

Vispoel, W. P., Wang, T., Torre, R. d. l., Bleiler, T., & Dings, J. (1992, April). How review options and administration modes influence scores on computerized vocabulary tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Wild, C. L., Durso, R., & Rubin, D. R. (1982). Effect of increased test-taking time on test scores by ethnic group, years out of school, and sex. Journal of Educational Measurement, 19, 19-28.

Wise, S. L. (1994). Understanding self-adapted testing: The perceived control hypothesis. Applied Measurement in Education, 7, 15-24.

Wise, S. L. (1996, April). A critical analysis of the arguments for and against item review in computerized adaptive testing. Paper presented at the annual meeting of the National Conference on Measurement in Education, New York, NY.

Wise, S. L., Barnes, L. B., Harvey, A. L., & Plake, B. S. (1989). Effects of computer anxiety and computer experience on the computer-based achievement test performance of college students. Applied Measurement in Education, 2, 235-241.

Wise, S. L., Roos, L. L., Leland, V. L., Oats, R. G., & McCrann, T. O. (1996). The development and validation of a scale measuring desire for control on examinations. Educational and Psychological Measurement, 56, 710-718.

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. Applied Measurement in Education, 8, 227-242.

Zara, A. R. (1992, April). An investigation of computerized adaptive testing for demographically-diverse candidates on the national registered nurse licensure examination. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.